Document-level Automatic MT Evaluation based on Discourse Representations

Jesús Giménez and **Lluís Màrquez**
TALP UPC
Barcelona, Spain
{jgimenez, lluism}@lsi.upc.edu

Elisabet Comelles and **Irene Castellón**
Universitat de Barcelona
Barcelona, Spain
{elicomelles, icastellon}@ub.edu

Victoria Arranz
ELDA/ELRA
Paris, France
arranz@elda.org

Abstract

This paper describes the joint submission of Universitat Politècnica de Catalunya and Universitat de Barcelona to the Metrics MaTr 2010 evaluation challenge, in collaboration with ELDA/ELRA. Our work is aimed at widening the scope of current automatic evaluation measures from sentence to document level. Preliminary experiments, based on an extension of the metrics by Giménez and Màrquez (2009) operating over discourse representations, are presented.

1 Introduction

Current automatic similarity measures for Machine Translation (MT) evaluation operate all, without exception, at the segment level. Translations are analyzed on a segment-by-segment[1] fashion, ignoring the text structure. Document and system scores are obtained using aggregate statistics over individual segments. This strategy presents the main disadvantage of ignoring cross-sentential/discursive phenomena.

In this work we suggest widening the scope of evaluation methods. We have defined genuine document-level measures which are able to exploit the structure of text to provide more informed evaluation scores. For that purpose we take advantage of two coincidental facts. First, test beds employed in recent MT evaluation campaigns include a document structure grouping sentences related to the same event, story or topic (Przybocki et al., 2008; Przybocki et al., 2009; Callison-Burch et al., 2009). Second, we count on automatic linguistic processors which provide very detailed discourse-level representations of text (Curran et al., 2007).

Discourse representations allow us to focus on relevant pieces of information, such as the agent (who), location (where), time (when), and theme (what), which may be spread all over the text. Counting on a means of discerning the events, the individuals taking part in each of them, and their role, is crucial to determine the semantic equivalence between a reference document and a candidate translation.

Moreover, the discourse analysis of a document is not a mere concatenation of the analyses of its individual sentences. There are some phenomena which may go beyond the scope of a sentence and can only be explained within the context of the whole document. For instance, in a newspaper article, facts and entities are progressively added to the discourse and then referred to anaphorically later on. The following extract from the development set illustrates the importance of such a phenomenon in the discourse analysis: '*Among the current or underlying crises in the Middle East, Rod Larsen mentioned the Arab-Israeli conflict and the Iranian nuclear portfolio, as well as the crisis between Lebanon and Syria. He stated: "All this leads us back to crucial values and opinions, which render the situation prone at any moment to getting out of control, more so than it was in past days."*'. The subject pronoun *"he"* works as an anaphoric pronoun whose antecedent is the proper noun *"Rod Larson"*. The anaphoric relation established between these two elements can only be identified by analyzing the text as a whole, thus considering the gender agreement between the third person singular masculine subject pronoun *"he"* and the masculine proper noun *"Rod Larson"*. However, if the two sentences were analyzed separately, the identification of this anaphoric relation would not be feasible due to the lack of connection between the two elements. Discourse representations allow us to trace links across sentences between the different facts and entities appearing in them. Therefore, providing an approach to the text more similar to that of

[1] A segment typically consists of one or two sentences.

Proceedings of the Joint 5th Workshop on Statistical Machine Translation and MetricsMATR, pages 333–338,
Uppsala, Sweden, 11-16 July 2010. ©2010 Association for Computational Linguistics

a human, which implies taking into account the whole text structure instead of considering each sentence separately.

The rest of the paper is organized as follows. Section 2 describes our evaluation methods and the linguistic theory upon which they are based. Experimental results are reported and discussed in Section 3. Section 4 presents the metric submitted to the evaluation challenge. Future work is outlined in Section 5.

As an additional result, document-level metrics generated in this study have been incorporated to the IQMT package for automatic MT evaluation[2].

2 Metric Description

This section provides a brief description of our approach. First, in Section 2.1, we describe the underlying theory and give examples on its capabilities. Then, in Section 2.2, we describe the associated similarity measures.

2.1 Discourse Representations

As previously mentioned in Section 1, a document has some features which need to be analyzed considering it as a whole instead of dividing it up into sentences. The anaphoric relation between a subject pronoun and a proper noun has already been exemplified. However, this is not the only anaphoric relation which can be found inside a text, there are some others which are worth mentioning:

- the connection between a possessive adjective and a proper noun or a subject pronoun, as exemplified in the sentences *"Maria bought a new sweater. Her new sweater is blue.*", where the possessive feminine adjective *"her"* refers to the proper noun *"Maria"*.

- the link between a demonstrative pronoun and its referent, which is exemplified in the sentences *"He developed a new theory on grammar. However, this is not the only theory he developed"*. In the second sentence, the demonstrative pronoun *"this"* refers back to the noun phrase *"new theory on grammar"* which occurs in the previous sentence.

- the relation between a main verb and an auxiliary verb in certain contexts, as illustrated in the following pair of sentences *"Would you like more sugar? Yes, I would"*. In this example, the auxiliary verb *"would"* used in the short answer substitutes the verb phrase *"would like"*.

In addition to anaphoric relations, other features need to be highlighted, such as the use of discourse markers which help to give cohesion to the text, link parts of a discourse and show the relations established between them. Below, some examples are given:

- "Moreover", "Furthermore", "In addition" indicate that the upcoming sentence adds more information.

- "However", "Nonetheless", "Nevertheless" show contrast with previous ideas.

- "Therefore", "As a result", "Consequently" show a cause and effect relation.

- "For instance", "For example" clarify or illustrate the previous idea.

It is worth noticing that anaphora, as well as discourse markers, are key features in the interface between syntax, semantics and pragmatics. Thus, when dealing with these phenomena at a text level we are not just looking separately at the different language levels, but we are trying to give a complete representation of both the surface and the deep structures of a text.

2.2 Definition of Similarity Measures

In this work, as a first proposal, instead of elaborating on novel similarity measures, we have borrowed and extended the Discourse Representation *(DR)* metrics defined by Giménez and Màrquez (2009). These metrics analyze similarities between automatic and reference translations by comparing their respective discourse representations over individual sentences.

For the discursive analysis of texts, DR metrics rely on the C&C Tools (Curran et al., 2007), specifically on the Boxer component (Bos, 2008). This software is based on the Discourse Representation Theory (DRT) by Kamp and Reyle (1993). DRT is a theoretical framework offering a representation language for the examination of contextually dependent meaning in discourse. A discourse is represented in a discourse representation structure (DRS), which is essentially a variation of first-order predicate calculus —its forms are pairs

[2] `http://www.lsi.upc.edu/~nlp/IQMT`

of first-order formulae and the free variables that occur in them.

DRSs are viewed as semantic trees, built through the application of two types of DRS conditions:

basic conditions: one-place properties (predicates), two-place properties (relations), named entities, time-expressions, cardinal expressions and equalities.

complex conditions: disjunction, implication, negation, question, and propositional attitude operations.

For instance, the DRS representation for the sentence "*Every man loves Mary.*" is as follows: $\exists y \; named(y, mary, per) \land (\forall x \; man(x) \rightarrow \exists z \; love(z) \land event(z) \land agent(z, x) \land patient(z, y))$. DR integrates three different kinds of metrics:

DR-STM These metrics are similar to the *Syntactic Tree Matching* metric defined by Liu and Gildea (2005), in this case applied to DRSs instead of constituent trees. All semantic subpaths in the candidate and reference trees are retrieved. The fraction of matching subpaths of a given length (l=4 in our experiments) is computed.

DR-$O_r(\star)$ Average lexical overlap between discourse representation structures of the same type. Overlap is measured according to the formulae and definitions by Giménez and Màrquez (2007).

DR-$O_{rp}(\star)$ Average morphosyntactic overlap, i.e., between grammatical categories –parts-of-speech– associated to lexical items, between discourse representation structures of the same type.

We have extended these metrics to operate at document level. For that purpose, instead of running the C&C Tools in a sentence-by-sentence fashion, we run them document by document. This is as simple as introducing a "<META>" tag at the beginning of each document to denote document boundaries[3].

3 Experimental Work

In this section, we analyze the behavior of the new DR metrics operating at document level with respect to their sentence-level counterparts.

3.1 Settings

We have used the 'mt06' part of the development set provided by the Metrics MaTr 2010 organization, which corresponds to a subset of 25 documents from the NIST 2006 Open MT Evaluation Campaign Arabic-to-English translation. The total number of segments is 249. The average number of segments per document is, thus, 9.96. The number of segments per document varies between 2 and 30. For the purpose of automatic evaluation, 4 human reference translations and automatic outputs by 8 different MT systems are available. In addition, we count on the results of a process of manual evaluation. Each translation segment was assessed by two judges. After independently and completely assessing the entire set, the judges reviewed their individual assessments together and settled on a single final score. Average system adequacy is 5.38.

In our experiments, metrics are evaluated in terms of their correlation with human assessments. We have computed Pearson, Spearman and Kendall correlation coefficients between metric scores and adequacy assessments. Document-level and system-level assessments have been obtained by averaging over segment-level assessments. We have computed correlation coefficients and confidence intervals applying bootstrap resampling at a 99% statistical significance (Efron and Tibshirani, 1986; Koehn, 2004). Since the cost of exhaustive resampling was prohibitive, we have limited to 1,000 resamplings. Confidence intervals, not shown in the tables, are in all cases lower than 10^{-3}.

3.2 Metric Performance

Table 1 shows correlation coefficients at the document level for several DR metric representatives, and their document-level counterparts (DR$_{doc}$). For the sake of comparison, the performance of the METEOR metric is also reported[4].

Contrary to our expectations, DR$_{doc}$ variants obtain lower levels of correlation than their DR

[3] Details on the advanced use of Boxer are available at `http://svn.ask.it.usyd.edu.au/trac/candc/wiki/BoxerComplex`.

[4] We have used METEOR version 1.0 with default parameters optimized by its developers over adequacy and fluency assessments. The METEOR metric is publicly available at `http://www.cs.cmu.edu/~alavie/METEOR/`

Metric	Pearson_ρ	Spearman_ρ	Kendall_τ
METEOR	**0.9182**	**0.8478**	**0.6728**
DR-$O_r(\star)$	0.8567	0.8061	0.6193
DR-$O_{rp}(\star)$	0.8286	0.7790	0.5875
DR-STM	0.7880	0.7468	0.5554
DR$_{doc}$-$O_r(\star)$	0.7936	0.7784	0.5875
DR$_{doc}$-$O_{rp}(\star)$	0.7219	0.6737	0.4929
DR$_{doc}$-STM	0.7553	0.7421	0.5458

Table 1: Meta-evaluation results at document level

Metric	Pearson_ρ	Spearman_ρ	Kendall_τ
METEOR	**0.9669**	0.9151	0.8533
DR-$O_r(\star)$	0.9100	0.6549	0.5764
DR-$O_{rp}(\star)$	0.9471	0.7918	0.7261
DR-STM	0.9295	0.7676	0.7165
DR$_{doc}$-$O_r(\star)$	0.9534	0.8434	0.7828
DR$_{doc}$-$O_{rp}(\star)$	0.9595	0.9101	0.8518
DR$_{doc}$-STM	**0.9676**	**0.9655**	**0.9272**
DR-$O_r(\star)'$	0.9836	0.9594	0.9296
DR-$O_{rp}(\star)'$	**0.9959**	**1.0000**	**1.0000**
DR-STM$'$	0.9933	0.9634	0.9307

Table 2: Meta-evaluation results at system level

counterparts. There are three different factors which could provide a possible explanation for this negative result. First, the C&C Tools, like any other automatic linguistic processor are not perfect. Parsing errors could be causing the metric to confer less informed scores. This is especially relevant taking into account that candidate translations are not always well-formed. Secondly, we argue that the way in which we have obtained document-level quality assessments, as an average of segment-level assessments, may be biasing the correlation. Thirdly, perhaps the similarity measures employed are not able to take advantage of the document-level features provided by the discourse analysis. In the following subsection we show some error analysis we have conducted by inspecting particular cases.

Table 2 shows correlation coefficients at system level. In the case of DR and DR$_{doc}$ metrics, system scores are computed by simple average over individual documents. Interestingly, in this case DR$_{doc}$ variants seem to obtain higher correlation than their DR counterparts. The improvement is especially substantial in terms of Spearman and Kendall coefficients, which do not consider absolute values but ranking positions. However, it could be the case that it was just an average effect. While DR metrics compute system scores as an average of segment scores, DR$_{doc}$ metrics average directly document scores. In order to clarify this result, we have modified DR metrics so as to compute system scores as an average of document scores (DR$'$ variants, the last three rows in the table). It can be observed that DR' variants outperform their DR$_{doc}$ counterparts, thus confirming our suspicion about the averaging effect.

3.3 Analysis

It is worth noting that DR$_{doc}$ metrics are able to detect and deal with several linguistic phenomena related to both syntax and semantics at sentence and document level. Below, several examples illustrating the potential of this metric are presented.

Control structures. Control structures (either subject or object control) are always a difficult issue as they mix both syntactic and semantic knowledge. In Example 1 a couple of control structures must be identified and DR$_{doc}$ metrics deal correctly with the argument structure of all the verbs involved. Thus, in the first part of the sentence, a subject control verb can be identified being *"the minister"* the agent of both verb forms *"go"* and *"say"*. On the other hand, in the quoted question, the verb *"invite"* works as an object control verb because its patient *"Chechen representatives"* is also the agent of the verb *visit*.

Example 1: *The minister went on to say, "What would Moscow say if we were to invite Chechen representatives to visit Jerusalem?"*

Anaphora and pronoun resolution. Whenever there is a pronoun whose antecedent is a named entity (NE), the metric identifies correctly its antecedent. This feature is highly valuable because a relationship between syntax and semantics is established. Moreover, when dealing with Semantic Roles the roles of Agent or Patient are given to the antecedents instead of the pronouns. Thus, in Example 2 the antecedent of the relative pronoun *"who"* is the NE *"Putin"* and the patient of the verb *"classified"* is also the NE *"Putin"* instead of the relative pronoun *"who"*.

Example 2: *Putin, who was not classified as his country Hamas as "terrorist organizations", recently said that the European Union is "a big mistake" if it decided to suspend financial aid to the Palestinians.*

Nevertheless, although Boxer was expected to deal with long-distance anaphoric relations beyond the sentence, after analyzing several cases, results show that it did not succeed in capturing this type of relations as shown in Example 3. In this example, the antecedent of the pronoun *"he"* in the second sentence is the NE *"Roberto Calderoli"* which appears in the first sentence. DR$_{doc}$ metrics should be capable of showing this connection. However, although the proper noun *"Roberto Calderoli"* is identified as a NE, it does not share the same reference as the third person singular pronoun *"he"*.

Example 3: *Roberto Calderoli does not intend to apologize. The newspaper Corriere Della Sera reported today, Saturday, that he said "I don't feel responsible for those deaths."*

4 Our Submission

Instead of participating with individual metrics, we have combined them by averaging their scores as described in (Giménez and Màrquez, 2008). This strategy has proven as an effective means of combining the scores conferred by different metrics (Callison-Burch et al., 2008; Callison-Burch et al., 2009). Metrics submitted are:

DR$_{doc}$ an arithmetic mean over a heuristically-defined set of DR$_{doc}$ metric variants, respectively computing lexical overlap, morphosyntactic overlap, and semantic tree matching (M = {*'DR$_{doc}$-$O_r(\star)$', 'DR$_{doc}$-$O_{rp}(\star)$', 'DR$_{doc}$-STM$_4$'*}). Since DR$_{doc}$ metrics do not operate over individual segments, we have assigned each segment the score of the document in which it is contained.

DR a measure analog to DR$_{doc}$ but using the default version of DR metrics operating at the segment level (M = {*'DR-$O_r(\star)$', 'DR-$O_{rp}(\star)$', 'DR-STM$_4$'*}).

ULC$_h$ an arithmetic mean over a heuristically-defined set of metrics operating at different linguistic levels, including lexical metrics, and measures of overlap between constituent parses, dependency parses, semantic roles, and discourse representations (M = {*'ROUGE$_W$', 'METEOR', 'DP-HWC$_r$', 'DP-$O_c(\star)$', 'DP-$O_l(\star)$', 'DP-$O_r(\star)$', 'CP-STM$_4$', 'SR-$O_r(\star)$', 'SR-O_{rv}', 'DR-$O_{rp}(\star)$'*}). This metric corresponds exactly to the metric submitted in our previous participation.

The performance of these metrics at the document and system levels is shown in Table 3.

5 Conclusions and Future Work

We have presented a modified version of the DR metrics by Giménez and Màrquez (2009) which, instead of limiting their scope to the segment level, are able to capture and exploit document-level features. However, results in terms of correlation with human assessments have not reported any improvement of these metrics over their sentence-level counterparts as document and system quality predictors. It must be clarified whether the problem is on the side of the linguistic tools, in the similarity measure, or in the way in which we have built document-level human assessments.

For future work, we plan to continue the error analysis to clarify why DR$_{doc}$ metrics do not outperform their DR counterparts at the document level, and how to improve their behavior. This

Metric	Document level			System level		
	Pearson$_\rho$	Spearman$_\rho$	Kendall$_\tau$	Pearson$_\rho$	Spearman$_\rho$	Kendall$_\tau$
ULC$_{DR}$	0.8418	0.8066	0.6135	0.9349	0.7936	0.7145
ULC$_{DRdoc}$	0.7739	0.7358	0.5474	0.9655	0.9062	0.8435
ULC$_h$	0.8963	0.8614	0.6848	0.9842	0.9088	0.8638

Table 3: Meta-evaluation results at document and system level for submitted metrics

may imply defining new metrics possibly using alternative linguistic processors. In addition, we plan to work on the identification and analysis of discourse markers. Finally, we plan to repeat this experiment over other test beds with document structure, such as those from the 2009 Workshop on Statistical Machine Translation shared task (Callison-Burch et al., 2009) and the 2009 NIST MT Evaluation Campaign (Przybocki et al., 2009). In the case that document-level assessments are not provided, we will also explore the possibility of producing them ourselves.

Acknowledgments

This work has been partially funded by the Spanish Government (projects OpenMT-2, TIN2009-14675-C03, and KNOW, TIN-2009-14715-C0403) and the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement numbers 247762 (FAUST project, FP7-ICT-2009-4-247762) and 247914 (MOLTO project, FP7-ICT-2009-4-247914). We are also thankful to anonymous reviewers for their comments and suggestions.

References

Johan Bos. 2008. Wide-coverage semantic analysis with boxer. In Johan Bos and Rodolfo Delmonte, editors, *Semantics in Text Processing. STEP 2008 Conference Proceedings*, Research in Computational Semantics, pages 277–286. College Publications.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2008. Further meta-evaluation of machine translation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 70–106.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Josh Schroeder. 2009. Findings of the 2009 Workshop on Statistical Machine Translation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 1–28.

James Curran, Stephen Clark, and Johan Bos. 2007. Linguistically motivated large-scale nlp with c&c and boxer. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 33–36.

Bradley Efron and Robert Tibshirani. 1986. Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy. *Statistical Science*, 1(1):54–77.

Jesús Giménez and Lluís Màrquez. 2007. Linguistic Features for Automatic Evaluation of Heterogeneous MT Systems. In *Proceedings of the ACL Workshop on Statistical Machine Translation*, pages 256–264.

Jesús Giménez and Lluís Màrquez. 2008. A Smorgasbord of Features for Automatic MT Evaluation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 195–198.

Jesús Giménez and Lluís Màrquez. 2009. On the Robustness of Syntactic and Semantic Features for Automatic MT Evaluation. In *Proceedings of the 4th Workshop on Statistical Machine Translation (EACL 2009)*.

Hans Kamp and Uwe Reyle. 1993. *From Discourse to Logic: An Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory*. Dordrecht: Kluwer.

Philipp Koehn. 2004. Statistical Significance Tests for Machine Translation Evaluation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 388–395.

Ding Liu and Daniel Gildea. 2005. Syntactic Features for Evaluation of Machine Translation. In *Proceedings of ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*, pages 25–32.

Mark Przybocki, Kay Peterson, and Sébastien Bronsart. 2008. NIST Metrics for Machine Translation 2008 Evaluation (MetricsMATR08). Technical report, National Institute of Standards and Technology.

Mark Przybocki, Kay Peterson, and Sébastien Bronsart. 2009. NIST Open Machine Translation 2009 Evaluation (MT09). Technical report, National Institute of Standards and Technology.